Connections and symbols II

AT1
Emmanuel Dupoux

Cogmaster, 2015

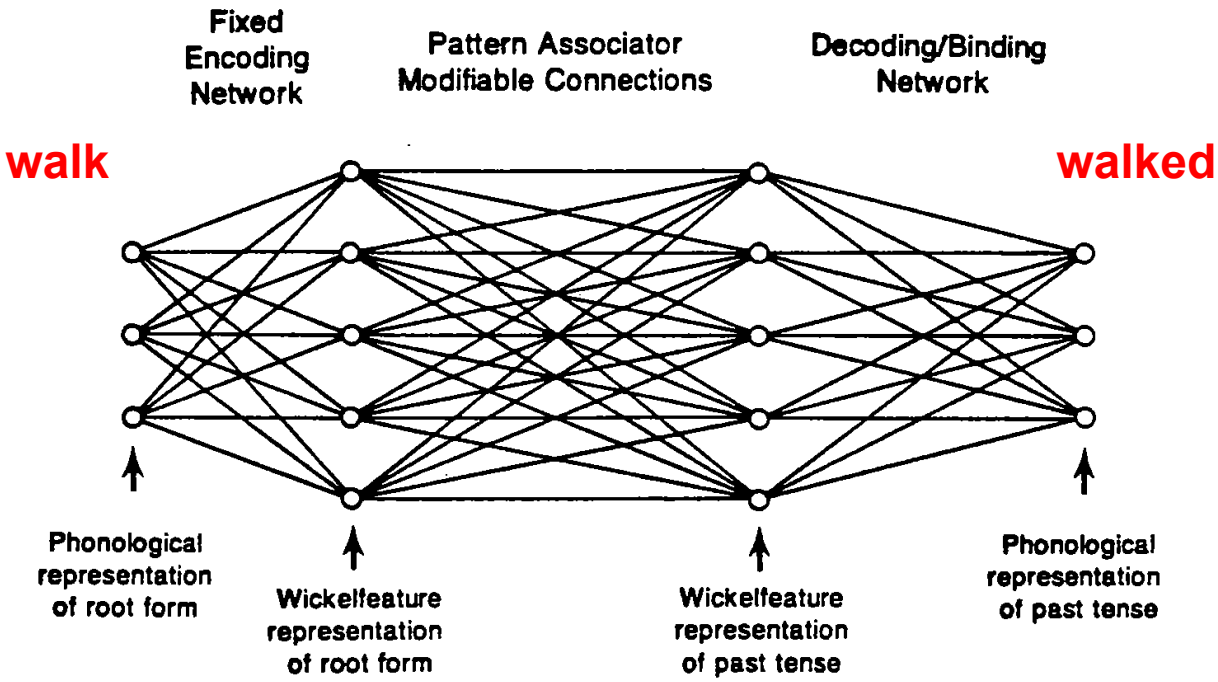
summary of preceding session

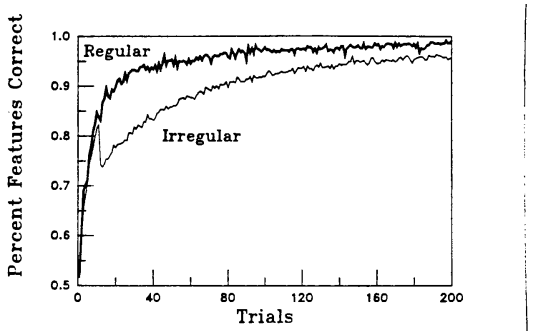
- how could a physical system ‘think’?
  - reasoning (mathematical, common sense)
  - performing complex cognitive tasks

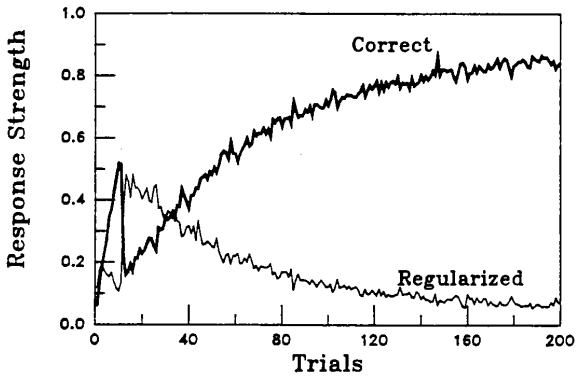
- philosophical issue: “computational reduction”
  - reduction of unboundedly complex behavior to the combination of simple ones
    - simple set of primitive processes
    - finite set of data types
    - a finite set of operations that combine the primitive processes to make more complex ones
- Scientific issue: precise (formal,quantitative) theories of cognition
  - what computational mechanisms underly complex behaviors (like language, reasoning, etc)?
    - Symbolic IA: (sequential & deterministic) computations with symbols and rules
      - eg: Turing machines, rewrite rules, finite state automata
    - Connexionist IA: (parallel & stochastic) computation with (continuous valued) neurone-like units
      - eg: Multilevel Perceptrons, Boltzman machines
- Technological issue: usable systems (IA)

McClelland & Rumelhart's (1986) Parallel Distributed Processing

- Cognition involves the spreading of activation, relaxation, statistical correlation.
- Represents a method for how symbolic systems might be implemented
  - Hypothesized that apparently symbolic processing is an emergent property of subsymbolic operations.
- Advantages
  - Fault tolerance & graceful degradation
  - Can be used to model learning
  - More naturally capture nonlinear relationships
  - Fuzzy information retrieval
  - bridges the gap with real neural processing







Rumelhart & McClelland (1986)

The **critique**: Pinker & Mehler (1988)

- Lachter & Bever: connectionist theories are a return to associationism (Chomsky vs Skinner revisited)
- Pinker & Prince: connectionist models of the capacity to derive the past tense of English verbs is inadequate
  - rules: wug → wugged
  - exceptions: put -> put, go->went, dig->dug
- Fodor & Pylyshyn: connectionnist theories are inadequate models of language and thought

Fodor & Pylyshyn

- Position of the problem: classical theories vs connectionnism
  - Agree:
    - both classical theories & connectionism are *representationalist*s (they assign some ‘meaning’ to the elements – symbols or nodes)
  - Disagree
    - classical theory encode structural relationships and processes (eg, constituents, variables, rules)
    - connectionnists only encore causal relationships and processes (x causes y to fire)
- Arguments against connectionnist systems: mental representation and processes are structure sensitive
  - combinatorial semantics
    - semantics of « J. loves M. » derived from semantics of « J. », « loves » and « M. »
  - productivity
    - the list of thoughts/sentences is not finite (I can construct new thoughts with old ones)
  - systematicity
    - I construct them in a systematic way
    - eg: « *x* loves M. » (where *x* can be any proper noun)
    - eg: If I can think « J. loves M. », I can think « M. loves J. »
  - recursivity & constituent structure:
    - If I can think « P. thinks that M. is nice » I can think « J thinks that P thinks that M is nice »

  -> connectionist systems have none of the above properties

Fodor & Pylyshyn (cont)

- Objections to symbolic/classical systems
  - rapidity of cognitive processes/neural speed
  - difficulty of pattern recognition/content based retrieval in conventional architectures
  - committed to rule vs exception dichotomy
  - inadaquate for intuitive /nonverbal behavior
  - acutely sensitive to damage/noise (vs graceful degradation)
  - storage in classical systems is passive
  - inadaquate account of gradual/frequency based application of rules
  - inadequate account of nondeterminism
  - no account of neuroscience

  → none of these arguments are valid or relevant

- CONCLUSIONS
  1. current connectionist theories are inadequate
  2. if they were to be made adequate they would be mere implementation of classical architecture

in brief

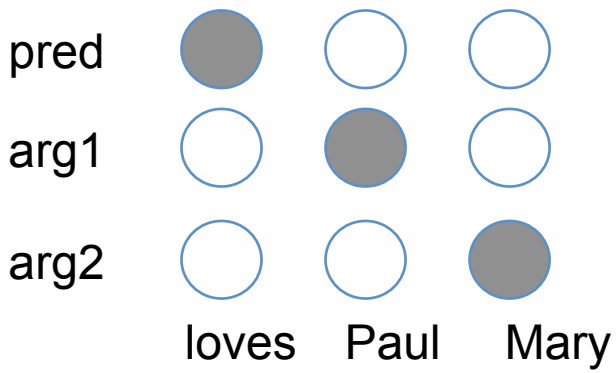
- The Fodor & Pylyshyn challenge:
  - (current) connectionist architectures fail to capture complex behaviors
  - (future) connectionist architectures are ‘mere’ implementation of symbolic architectures

- Structure of Smolensky
  - representing structures by fillers and roles
    - examples: trees, lists, etc
  - tensor products and filler/role binding (definition)
    - local, semilocal and distributed
  - unbinding (exact and self adressed)
  - capacity and graceful saturation
  - continuous and infinite structures
  - binding and unbinding networks
  - analogy between binding units and hebb weights
  - example of a stack
  - structured roles

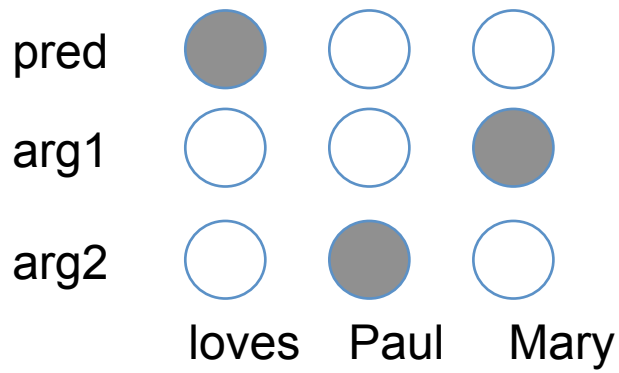
the Smolensky response: tensor products

Paul loves Mary-> loves(Paul,Mary)
pred=loves, arg1=Paul, arg2=Mary
pred*loves+arg1*Paul+arg2*Mary

Paul loves Mary
Mary is loved by Paul



Mary loves Paul
Paul is loved by Mary



→Problem of tensor product representations: exponential with sentence complexity

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structure in connectionnist systems. Artificial Intelligence, 46, 159-216.

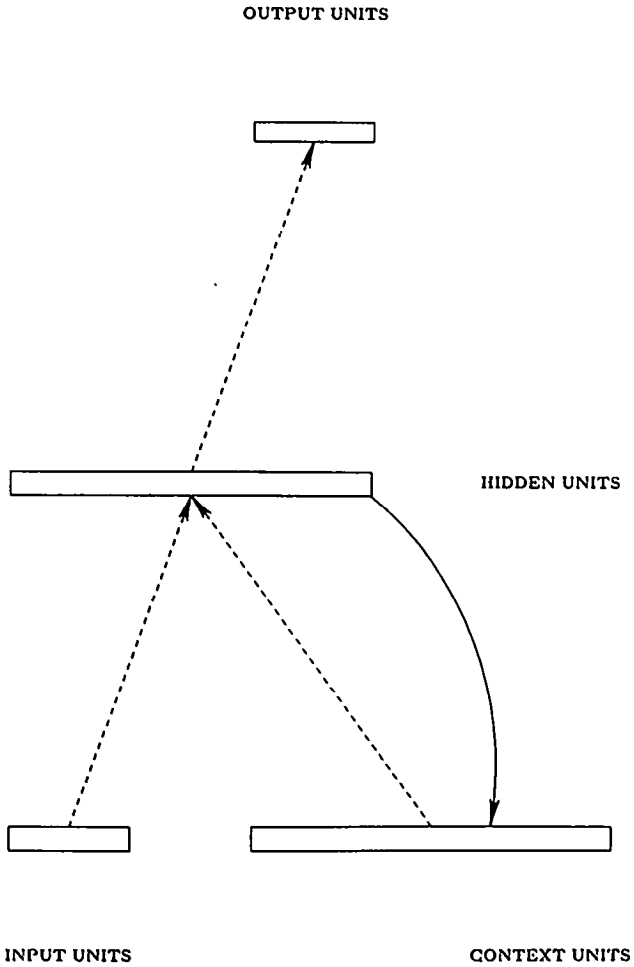
- extensions:
  - implementation of a phonological theory (Optimality Theory) in a tensor product network with energy relaxation
    - see the Harmonic Mind (Smolensky & Legendre, 2012)
  - Escaping the explosion in nb of neurons: holographic reduced representations
    - define A * B as an operation that preserves the dimensions (eg xor, circular convolution)

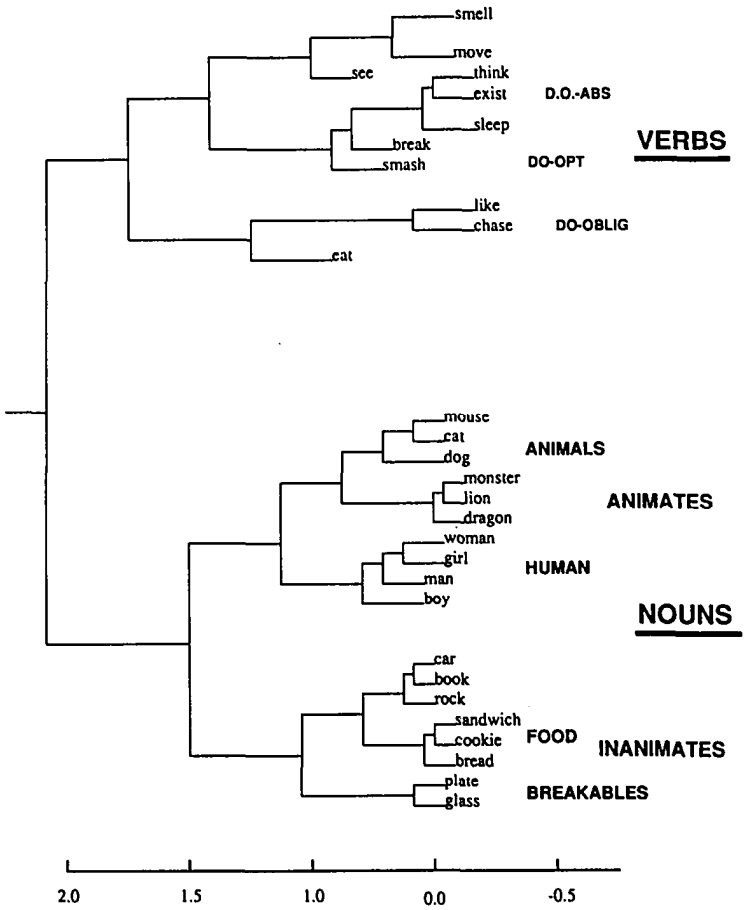
Plate (1995)

Elman

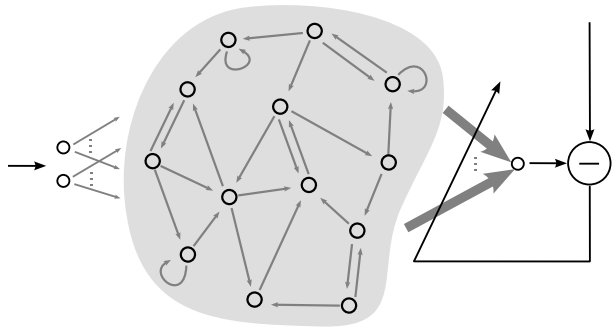
- structure of the paper
  - representing time
  - SRN architecture
  - xor through time
  - badiiguuu
  - word segmentation (15 words)
  - part of speech (13 categories, 29 words, 15 sentence templates)

Backprop applied





- extensions of Elman's SRN
  - computational capacity of SRN
    - Hyötyniemi, Heikki. "Turing machines are recurrent neural networks." Proceedings of step 96 (1996).
    - Servan-Schreiber, D., Cleeremans, A., & McClelland, J.L. (1988). Encoding sequential structure in simple recurrent networks (CMU Tech. Rep. No. CMU-CS-88-183). Pittsburgh, PA: Carnegie-Mellon University, Computer Science Department.
    - Lawrence, S., Giles, C. L., & Fong, S. (2000). Natural language grammatical inference with recurrent neural networks. IEEE Transactions on Knowledge and Data Engineering , 12(1), 126–140.
    - Pollack, J. B. (1991). The induction of dynamical recognizers. Machine Learning , 7(2–3), 227–252. R
    - Rodriguez, P. (2001). Simple recurrent networks learn context-free and context-sensitive languages by counting. Neural Computation, 13(9).
  - reservoir computing
    - http://reservoir-computing.org
    - Jaeger H (2007) Echo state network. Scholarpedia 2(9):2330. http://www.scholarpedia.org/article/Echo_state_network



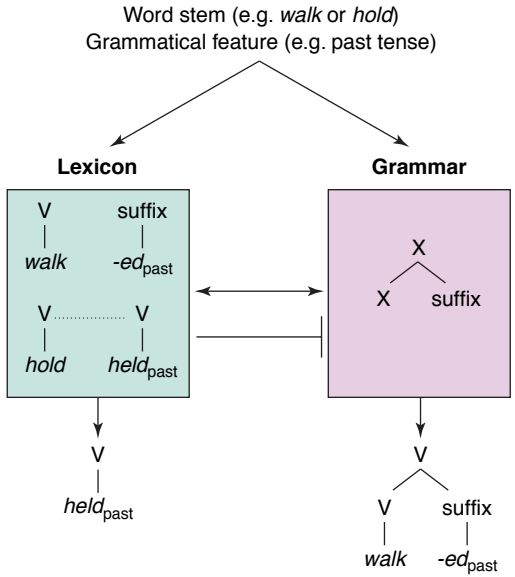
Conclusions

- What about the F&P Challenge?
  - tensor products are an interesting implementation/alternative to symbolic systems
  - recurrent networks could also be an alternative, but much less understood

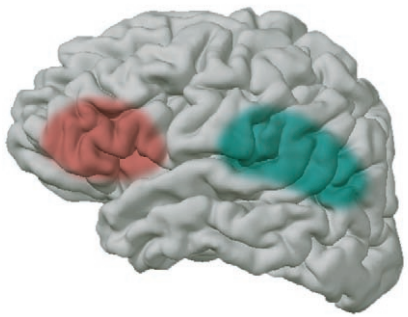
- The hidden debate
  - innate vs learner structures (to be continued…)

Conclusions

- empirical impact of the debate
  - past tense in English
    - rule: play->played, fax->faxed
    - exceptions: sing->sang, put->put
    - Pinker & Prince (1988)
    - procedural vs declarative memory (Ullman et al, 1997; Pinker & Ullman, 2002)



	Lexicon	Grammar
Used for:	roots, idioms, irregulars, some regulars	phrases, sentences, any regular form
Form of computation:	lookup, association	combination, unification
Subdivision of:	declarative memory	procedural system
Associated with:	words, facts	rules, skills
Principal substrate:	temporo-parietal cortex	frontal cortex, basal ganglia



Pinker, S. & Prince, A. (1988) On language and connectionism *Cognition*, 28, 73-193.
Ullman MT, Corkin S, et al. (1997). A neural dissociation within language: *Journal of Cognitive Neuroscience*, 9: 266–276.
Pinker, S. & Ullman, M. (2002) The past and future of the past tense. *Trends in Cognitive Science*, 6, 456-463.

Conclusions

- empirical impact of the debate (cont)
  - statistical learning vs algebraic learning in infants
    - Saffran et al, (1996), Marcus et al, (1999), Pena et al (2002)

Saffran, J., Aslin,R., Newport, E. (1996). *Science*
Marcus, G.F., Vijayan, S., Bandi Rao, S., Vishton, P. M. (1999). *Science*
Peña, M., Bonatti, L., Nespor, M., Mehler J. (2002). *Science*

  - exemplar-based versus abstract representations
    - object recognition (Biederman & Gerhardstein, 1993), face recognition, speech recognition (eg Goldinger, 1988; Johnson 1997, Pierrehumbert 2001)

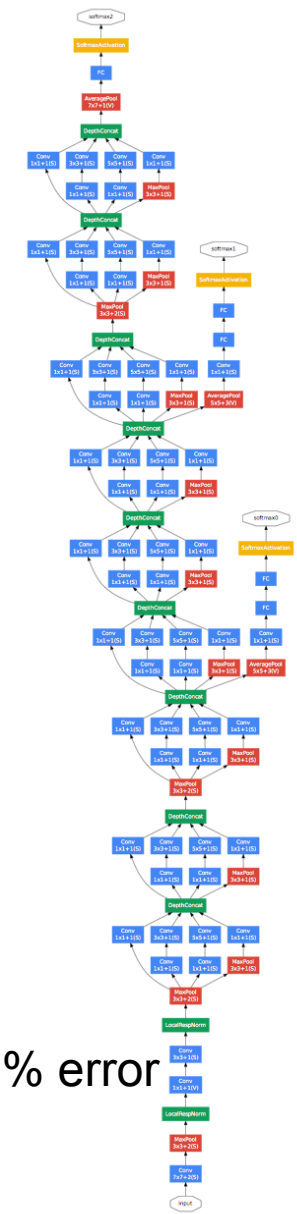
Biederman, I., & Gerhardstein, P. C. (1993). Journal of Experimental Psychology: Human Perception and Performance
Johnson, K. (1997). In K. Johnson & J.W. Mullennix (eds.), Talker Variability in Speech Processing,
Pierrehumbert, J. (2001). In J. Bybee and P. Hopper (eds.), Frequency and the Emergence of Linguistic Structure
Goldinger, S.D. (1998). *Psychological Review* 105:251-279.

Current trends

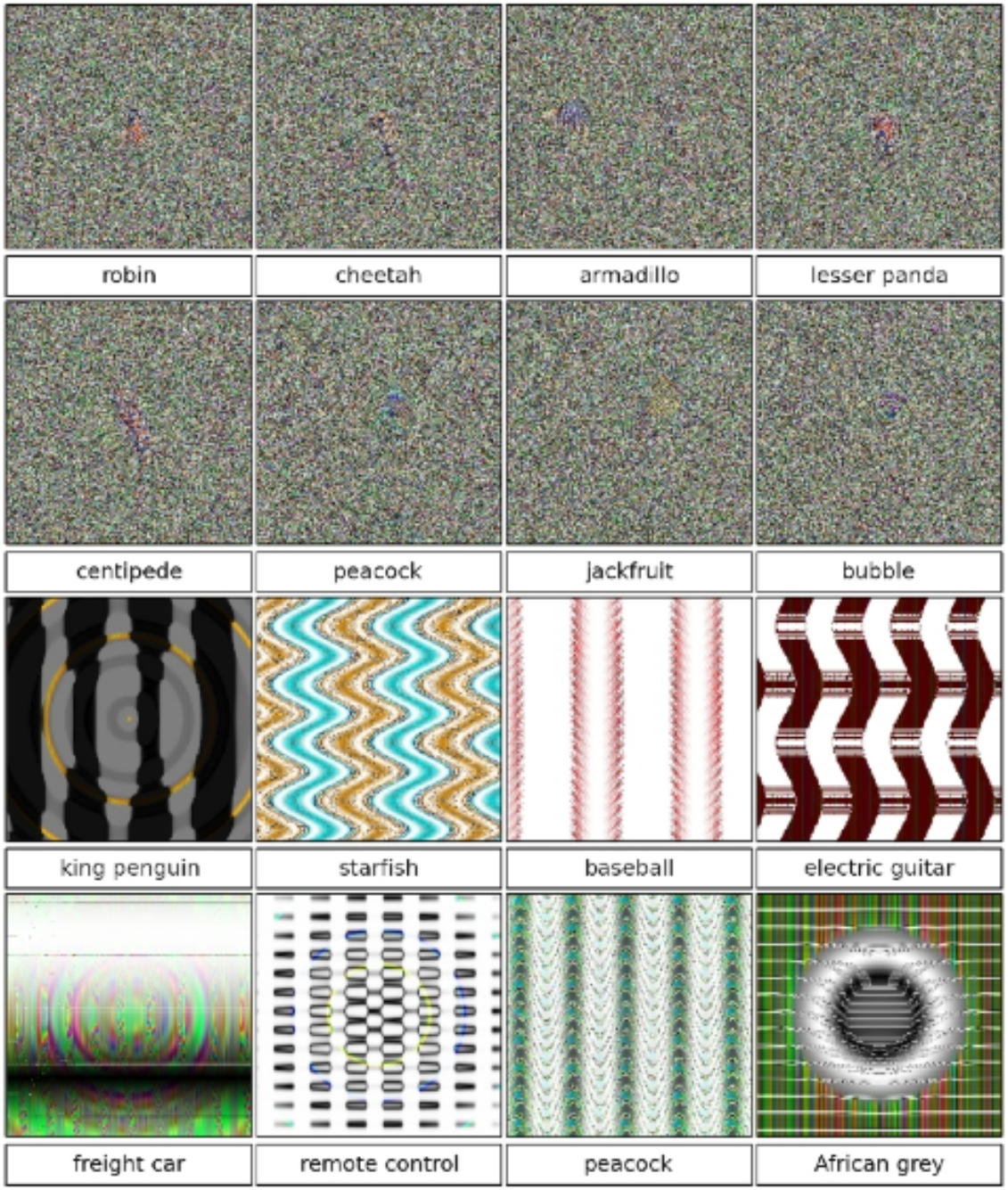
- very deep hierarchical networks
  - trained in a heavily supervised fashion
  - but, seem to correlate with human performance and monkey IT neurons

Yamins et al (2014) *Proc. Natl. Acad. Sci. U.S.A.*

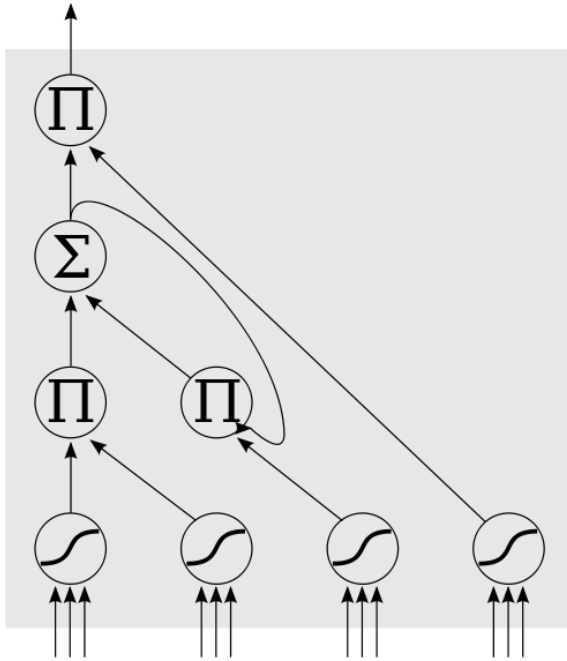
1000 categories
100.000 test
1M training
Google net: 6.67% error
Humans: 5%



Google net,
Szegedy et al, 2014


robin
cheetah
armadillo
lesser panda
centipede
peacock
jackfruit
bubble
king penguin
starfish
baseball
electric guitar
freight car
remote control
peacock
African grey

Following up on recurrent networks



- LSTM

  (long short term memory)

  Hochreiter and J. Schmidhuber (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

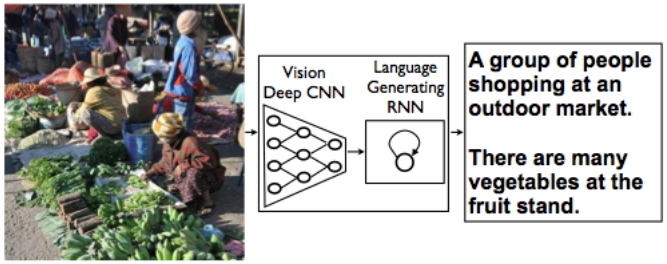
- replaces HMM for language models

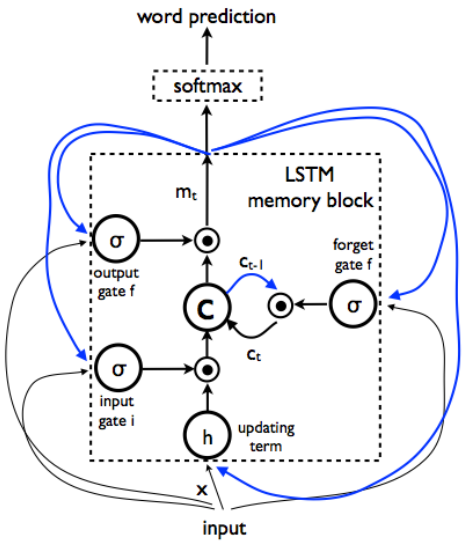
  Graves, Mohamed, Hinton (2013). Speech Recognition with Deep Recurrent Neural Networks. *ICASSP.*

- text generation

  Sutskever, Martens, Hinton (2011) Generating text with recurrent neural networks. ICML.

Image description





Vinyals et al (2014)
see also Donahue et al (2014)

- playing with RNNs
  - http://karpathy.github.io/2015/05/21/rnn-effectiveness/ (the unreasonable effectiveness of RNNs)

- memory networks
  - Weston Chopra & Bordes (2015)
  - question answering

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

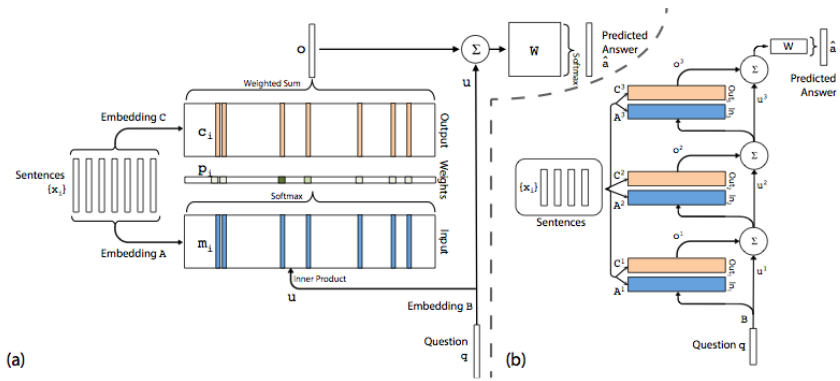
Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

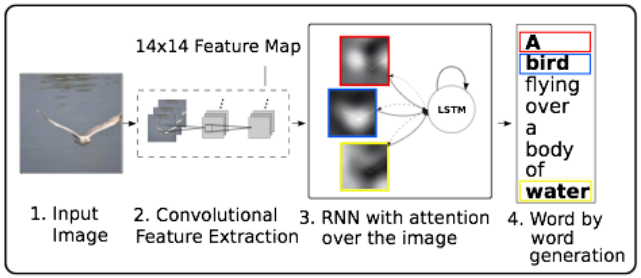
Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.



- attention networks
  - Xu, Ba, .. Bengio 2015



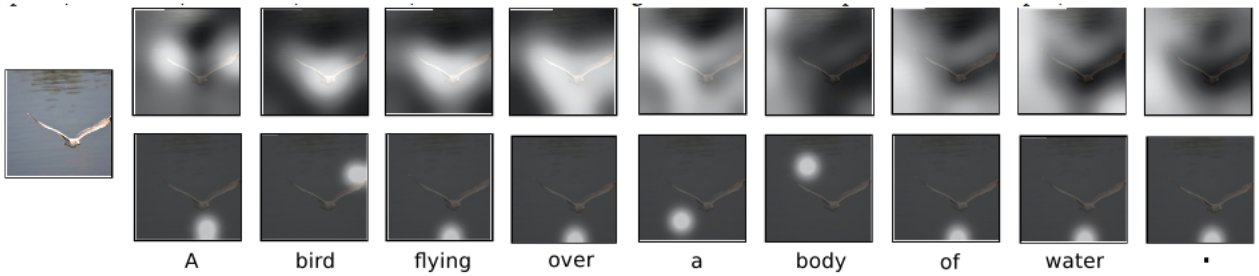
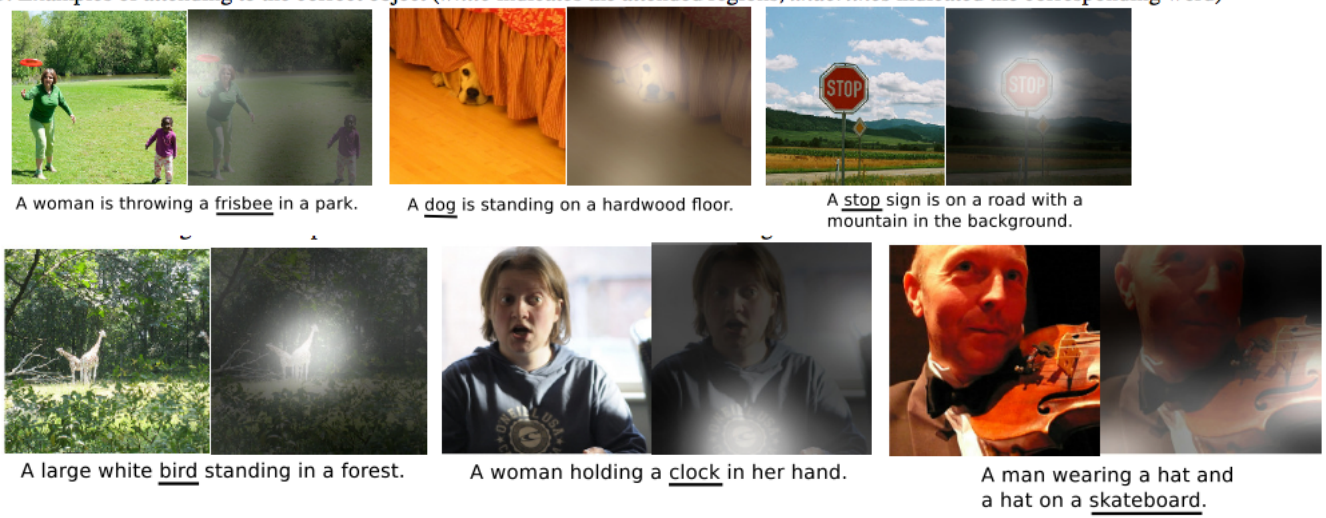
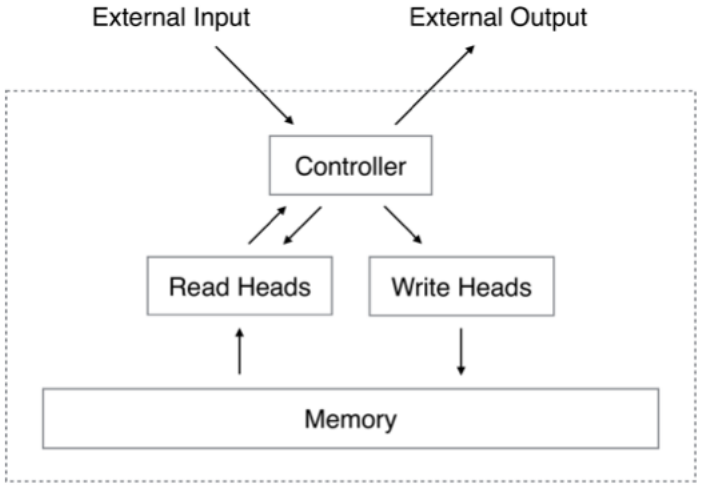
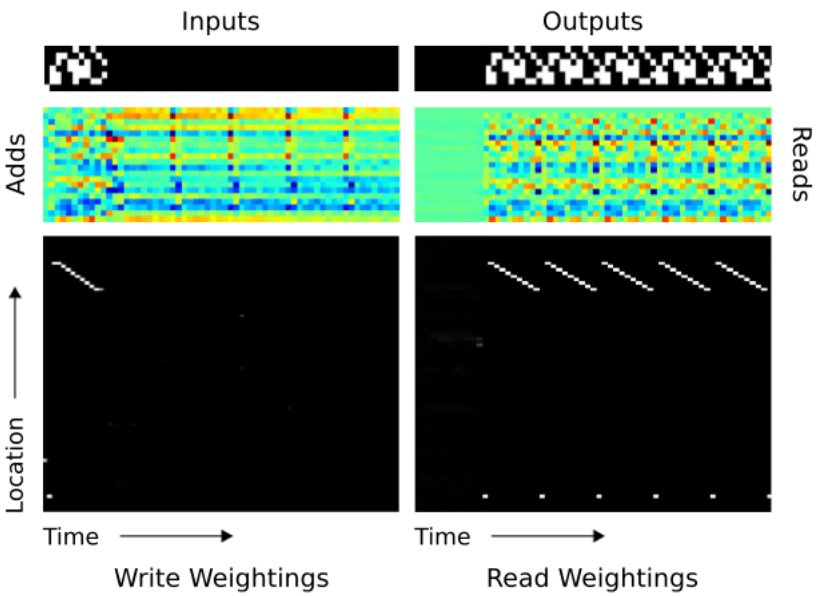


Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



- Neural Turing Machine
  - learns copy, sort





Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. *arXiv Preprint:1410.5401*.

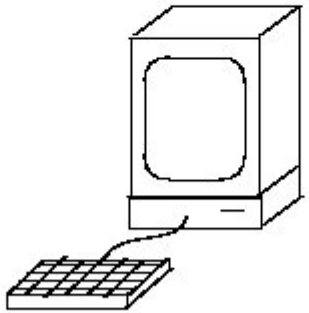
Evolutions of the symbolic systems

- probabilistic /bayesian frameworks

→ symbolic systems that learn

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems. San Mateo, CA: Morgan Kaufmann.*

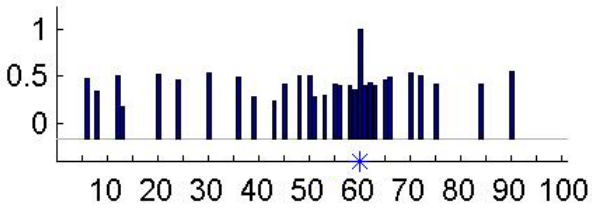
Le jeu des nombres



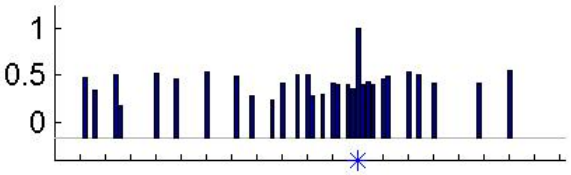
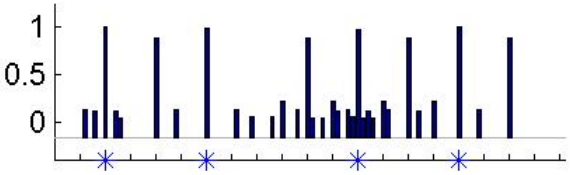
- input: nombre entre 1 et 100
- output: “oui” ou “non”

- Tâche d’induction:
  - Observer quelques exemples de ‘oui’
  - Juger si des nouveaux nombres auraient pu être “oui” ou “non”.

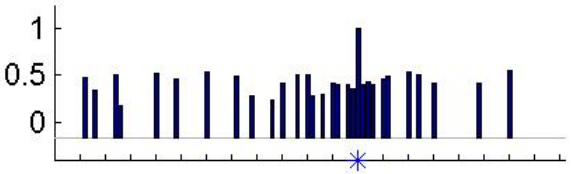
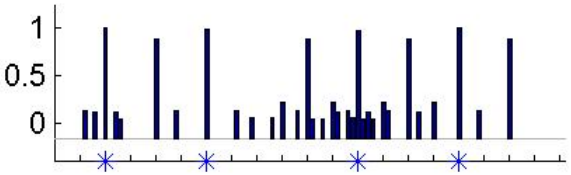
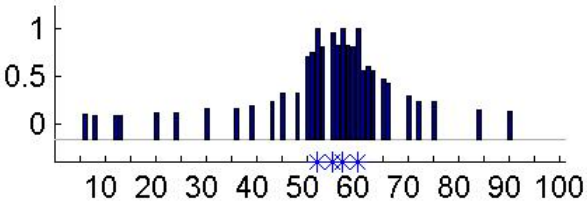
Le jeu des nombres

Exemples de “oui”	Jugements de généralisation (N = 20)
60	 Similarité diffuse

Le jeu des nombres

Exemples de “oui”	Jugements de généralisation (N = 20)	
60		Similarité diffuse
60 80 10 30		Règle: “multiples de 10”

Le jeu des nombres

Exemples de “oui”	Jugements de généralisation (N = 20)	
60		Similarité diffuse
60 80 10 30		Règle: “multiples de 10”
60 52 57 55		Similarité focalisée: nombres proches de 50-60

Modèle Bayesien

- *H*: Espace d'hypothèse des concepts possibles
  - h_1 = {2, 4, 6, 8, 10, 12, …, 96, 98, 100} ("nombres pairs")
  - h_2 = {10, 20, 30, 40, …, 90, 100} ("multiples de 10")
  - h_3 = {2, 4, 8, 16, 32, 64} ("puissances de 2")
  - h_4 = {50, 51, 52, …, 59, 60} ("nombres entre 50 et 60")
  - . . .

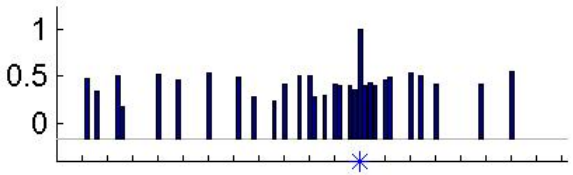
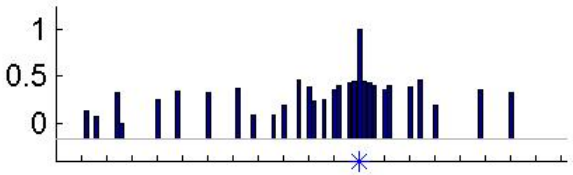
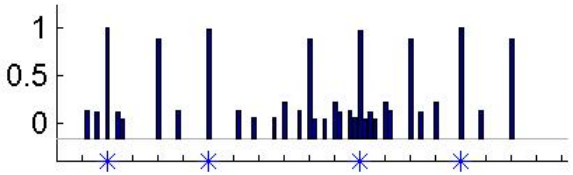
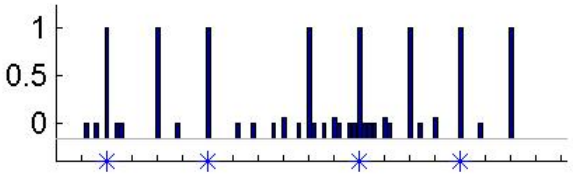
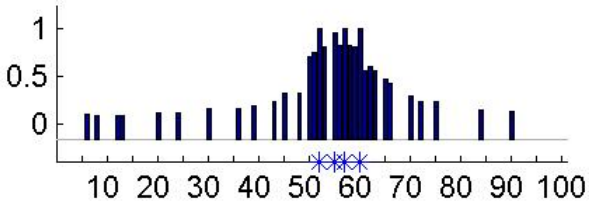
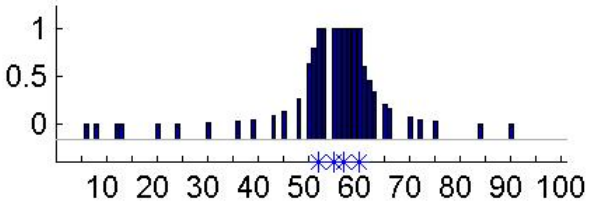
Modèle Bayesien

- *H*: Espace d'hypothèse des concepts possibles
  - h_1 = {2, 4, 6, 8, 10, 12, …, 96, 98, 100} ("nombres pairs")
  - h_2 = {10, 20, 30, 40, …, 90, 100} ("multiples de 10")
  - h_3 = {2, 4, 8, 16, 32, 64} ("puissances de 2")
  - h_4 = {50, 51, 52, …, 59, 60} ("nombres entre 50 et 60")
  - . . .
- $X = \{x_1, \ldots, x_n\}$: *n* exemples du concept *C*.
- Evaluer l'hypothèse h étant donné X:

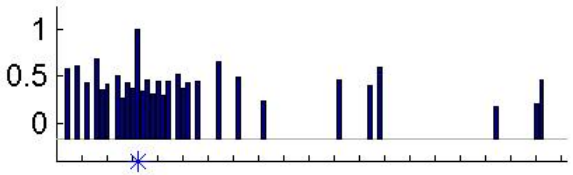
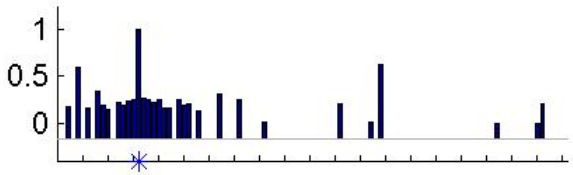
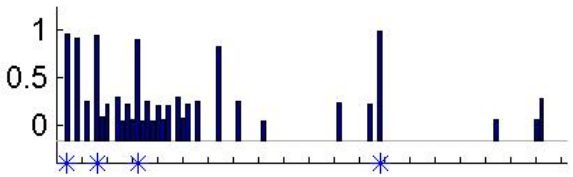
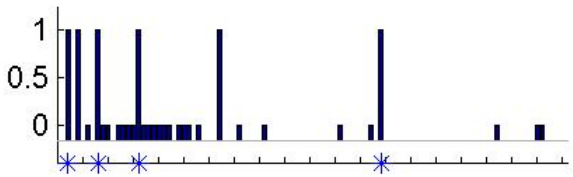
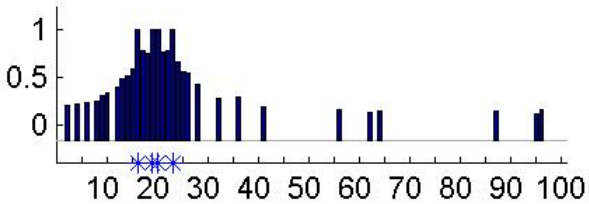
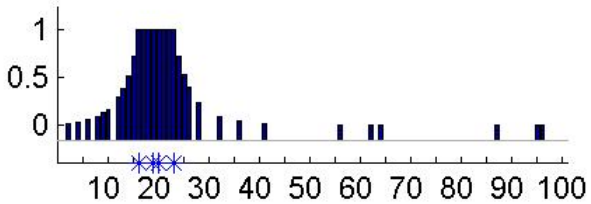
$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{\sum_{h' \in H} p(X \mid h')\, p(h')}$$

  - $p(h)$ ["prior"]: biais à priori
  - $p(X|h)$ ["likelihood"]: information statistique dans les exemples
  - $p(h|X)$ ["posterior"]: degré de confiance que *h* est l'extension de *C*.

Comparaison modèle-humain

Exemples de “oui”	Jugements (N = 20)	Modèle Bayesien (r = 0.96)
60		
60 80 10 30		
60 52 57 55		

Comparaison modèle-humain

Exemples de “oui”	Jugements (N = 20)	Modèle Bayesien (r = 0.93)
16		
16 8 2 64		
16 23 19 20		

from Josh Tennenbaum (MIT)

- Modèles linguistiques probabilistes/Natural Language processing
  - Automates finis stochastiques
  - Grammaires Context Free Probabilistes

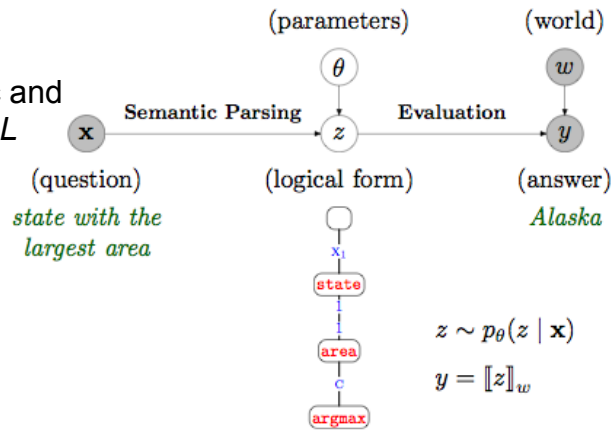
Utterance : you have another cookie

Candidate Meanings $\begin{cases} have(you, another(x, cookie(x))) \\ eat(you, your(x, cake(x))) \\ want(i, another(x, cookie(x))) \end{cases}$

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall.

- Learning of a syntactic/semantic parser
  - from pairs of sentence-meaning candidates

Kwiatkowski, T., Goldwater, S., et al (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. *EACL*



  - from pairs of questions and answers (plus a database of facts)

Liang, P., Jordan, M., I., & Klein, D. (2011). Learning Dependency-Based Compositional Semantics. î

Not covered here

- other connectionnist architectures
  - Kohonen's maps (competitive learning) (Kohonen, 1982)
  - Adaptive Resonance Theory (Grossberg, 1976)
  - Reinforcement learning (Barto, Sutton, Anderson, 1983)
- other computational frameworks
  - more Probabilistic/Bayesian frameworks
  - Predictive Coding/Free Energy (Friston 2009)